Prerequisites For Shallow-Transfer Machine Translation Of Mordvin Languages : Language Documentation With A Purpose

Rueter, Jack

þÿ 6 5 2 A : : = A B 8 B C B : > < ? L N B 5 @ = K E 8 A A ; 5 4 > 2 0 = 8 9
2020

Rueter , J & Hämäläinen , M 2020 , Prerequisites For Shallow-Transfer Machine Translation
þÿ O f M o r d v i n L a n g u a g e s : L a n g u a g e D o c u m e n t a t i o n W i t h A P u r p o s e . i n
þÿ 5 6 4 C = 0 @ > 4 = > 3 > > 1 @ 0 7 > 2 0 B 5 ; L = > 3 > A 0 ; > = 0 . 6 5 2 A : : = A B 8 B C B : > < ?
þÿ 8 A A ; 5 4 > 2 0 = 8 9 , I z h e v s k , p p . 1 8 - 2 9 .

http://hdl.handle.net/10138/325962

УДК 811.511.152’322=111(045)

Jack Rueter
PhD, researcher,
Finland, Helsinki, University of Helsinki

Mika Hämäläinen
PhD, researcher,
Finland, Helsinki, University of Helsinki

PREREQUISITES FOR SHALLOW-TRANSFER MACHINE TRANSLATION OF MORDVIN LANGUAGES: LANGUAGE DOCUMENTATION WITH A PURPOSE

В данной статье представлены текущие лексические, морфологические, синтаксические и основанные на правилах машинные переводы для эрзянского и мокшанского языков, которые можно и нужно использовать при разработке дорожной карты мордовских лингвистических исследований. Мы стремимся проиллюстрировать и обрисовать начальные типы проблем, с которыми придется столкнуться при построении системы поверхностного машинного перевода на основе Apertium для форм мордовских языков. Мы указываем ориентиры в мордовском языкознании и других разделах уралистики в качестве отправной точки для определения лингвистических исследований со средствами измерения собственного прогресса и разработки дорожной карты для дальнейших исследований.

Ключевые слова: *эрзянский, мокшанский, уральский, поверхностный машинный перевод, измеримые языковые исследования, измеримое языковое расстояние, конечная морфология, универсальные зависимости*.

Introduction

In this paper, we describe the current state of the on-going efforts in building a rule-based machine translation system for the Erzya-Moksha language pair[10]. In addition, we seek to identify possible ways to extend this work in such a way that the system can be built fast and with as much reuse as possible from existing tools and solutions.

There is a need to establish means to measure and develop a roadmap for Mordvin linguistic research. This research can and should include literary languages as well as dialect and historical language materials. By addressing normative and non-normative language forms with reference to temporal and spatial dimensions we can begin to fill the gaps in our knowledge of Mordvin language development and hypothetical proto forms. Writers and their publications can be associated with their geographical backgrounds. The spread of morphology, lexica, expressions and syntactic patterns within a language can be traced to their sources. Isoglosses can also be superimposed upon the entire group of closely related language forms. And finally, individual locales can be compared across language boundaries.

Linguistic descriptions of the Mordvin language forms Erzya, Moksha and Shoksha [cf. Rueter & Erina 2018] are often laboratory in character. When done by non-native speakers, the elementary purpose of language is often left out – communication. Non-native research typically attempts to treat the languages either as one conglomerate Mordvin or as a pool of knowledge to glean for typological features [Zaicz 1995; Gheno 1995; Hamari 2007; Keresztes 2011]. Even when research is conducted by native speakers on the individual language level, not enough attention is paid to the inspection of transferable knowledge; each language is seen as a masterpiece in its own right with little if any energy devoted to translation, i.e. information transfer [cf. Evsev′ev 1928/29; EKM 2000; MKM 2000; MKS 2008; EKS 2011].

10 Code available: https://github.com/apertium/apertium-myv-mdf

Related Work

While there has been several machine learning attempts to model machine translation in a low-resourced scenario, these approaches do not deal with endangered languages. For example, Kocmi & Bojar (2018) experiment with neural machine translation on big European languages and Arabic, Hämäläinen & Alnajjar (2019a) propose a template-based neural approach for Finnish to English, Ngo et al. (2020) treat Vietnamese as a low-resourced language. As these approaches are based on neural networks, they are data intensive to such a degree that simply is not possible in the endangered language scenario. For this reason, rule-based systems are more suitable for tackling the task.

One line of rule-based work focuses on modelling morphology of a language. Rule-based methods have been used for majority languages such as Finnish [Pirinen, 2015] and Russian [Reynolds, 2016]. Rule-based morphology has also been developed for many endangered languages such as the Mordvinic languages [Rueter et al., 2020], Plains Cree [Snoek et al., 2014], Skolt Sami [Rueter & Hämäläinen, 2020], Tatar [Salimzyanov et al., 2013] and Kven [Trosterud et al., 2017]. Rule-based morphology can also play an important role in higher level natural language generation tasks [Hämäläinen & Rueter, 2018; Hämäläinen & Alnajjar, 2019b].

Syntax and disambiguation have also been tackled with rule-based methods for several endangered languages [Uí Dhonnchadha & van Genabith 2006; Trosterud, 2009]. More recently, there have been efforts for using rule-based methods together with neural networks to achieve the same goal [Ens et al., 2019; Hämäläinen & Wiechetek, 2020]. A variety of dictionary building methods has also emerged for language documentation [Garrett, 2018; Rueter & Hämäläinen, 2017]. These tools provide a valuable resource for rule-based language modeling.

Apertium [Forcada et al., 2011] is a toolbox for developing rule-based machine translation systems. It uses other rule-based tools such as morphological generators and analyzers to conduct machine translation between a language pair. Apertium has been used successfully for many languages such as Finnish [Pirinen, 2019], Kazakh [Sundetova et al., 2014] Tatar and Bashkir [Tyers et al., 2012]. As its strength has been proven with many other languages, it is only natural that our work would be based on Apertium.

Treebanks provide a valuable resource for both linguistic and language technology research. The work on Komi-Permyak [Rueter, Partanen & Ponomareva, 2020] and Komi-Zyrian [Partanen et al., 2018] treebanks have been coordinated so that they also contain parallel sentences. Additionally, the tagging schemes are aligned with one another. Currently there is no Udmurt treebank, but we believe its emergence is only a matter of time.

Shallow-Transfer Machine Translation

Shallow-transfer machine translation can be most effective for translation between closely related languages (see [Corbí Bellot et al., 2005]). The idea behind this is that vocabulary, morphological categories and syntactic constructions have extensive reflexes in both the source and target languages. For people familiar with Erzya and Moksha studies [Rueter, Hämäläinen & Partanen, 2020].

In order to achieve low-transfer machine translation, a bilingual vocabulary is needed between the language pair. This can either be done manually or by deriving such a vocabulary automatically from multilingual dictionaries (see [Hämäläinen et al., 2018]). A recent development for manually building such dictionaries for Apertium has been the introduction of a web-based tool [Alnajjar et al., 2020] that makes it possible for non-technical people to contribute in building such resources.

Research in Erzya and Moksha can be done with the help of computer technology. This means that there are open-source descriptions of morphology, vocabulary, syntax, and translation research that can be found and are available on-line.

Since 1998, Rueter has been working on the finite-state description of the Erzyan and Moksha languages. Although most of the application of this work have not appeared until the last decade in the form of vocabularies and morphology, on the one hand (see [Rueter 2014; Moshagen et al 2014; Hämäläinen 2019]) and the description of Universal Dependencies, on the other [Rueter & Tyers 2018] and the task of improving their harmonization [Partanen & Rueter 2019]. With this in mind, it becomes possible to evaluate language research studies with the help of a rule-based translation system - in the Apertium project, based on shallow transfer.

The concept of applying shallow-transfer machine translation to measurement of linguistic diversity is relatively new. It involves combining three areas of language research into one project. As such it requires correlating parallel understanding and research of vocabulary, morphology and syntax in one formal field. This implies going beyond the achievements of comparative linguistics, thanks to which we have an understanding of etymologically related vocabulary and morphology. It also challenges earlier concepts of synonymy by introducing the requirement for the transfer of synonymous information from one linguistic community to another.

So what exactly is shallow-transfer rule-based translation and what does it provide as an engine for dialect and language research. In principle, the shallow transfer system will transfer the Erzya word *psakazo* 'his / her cat (nominative singular)' to the Moksha equivalent *katots* 'his / her cat (nominative singular)' in a three-step process. First, the word *psakazo* is analyzed with an FST (finite-state transducer) based morphological analyzer as *psaka*+N+Sg+Nom+PxSg3. Second, the Erzya noun *psaka* 'cat' is aligned with the corresponding Moksha word *kata* 'cat' using a vocabulary. Finally, the Moksha word *kata* 'cat' is generated with the equivalent morpho-semantic information, i.e. *kata*+N+Sg+Nom+PxSg3, which results in *katots* 'his / her cat' when inflected with an FST based morphological generator. When studying native speech, the Erzya word *psaka* has synonyms in *katka* and *pisaj*, and in Moksha there is at least a false friend for the Erzya, i.e. in Moksha one word for 'cat' is *kiska* (whereas *kiska* in Erzya means 'dog').

At the syntactic level, this same three-step process should be utilized for obtaining single readings or analyses for word forms in context. Without context, four distinct Moksha noun forms can be identified for possible syntactic functions (see examples 1.a-d), where the category of number is combined with the cases nominative and genitive and first person singular possession marker.

(1.a) *katoźe*: *kata*+N.Sg.Nom.PxSg1
(1.b) *katoźeń*: *kata*+N.Sg.Gen.PxSg1
(1.c) *katońe*: *kata*+N.Pl+Nom.PxSg1
(1.d) *katońeń*: *kata*+N.Pl.Gen.PxSg1

When the same morpho-semantic information is transferred to Erzya, however, what Moksha renders as four becomes two or possibly one singular form (see examples 2.a-d).

(2.a) *psaka*+N.Sg.Nom.PxSg1: *psakam*
(2.b) *psaka*+N.Sg.Gen.PxSg1: *psakam*, *psakan*
(2.c) *psaka*+N.Pl+Nom.PxSg1: *psakan*, *psakam*
(2.d) *psaka*+N.Pl.Gen.PxSg1: *psakan*, *psakam*

The only information we can obtain without context is that the form *psakan* in <*n*> cannot indicate nominative singular. Hence, a laboratory translation can, indeed, be made from Moksha to Erzya, on the one hand, resulting in the loss of distinctions in the categories of number and case if no context is present to retain the information. A reverse translation, on the other hand, will be useless without a context, as the single form *psakam* 'my cat' potentially introduces four possibilities for generation => *kata*+N.Sg.Nom.PxSg1, *kata*+N.Pl.Nom.PxSg1, *kata*+N.Sg.Gen.PxSg1, *kata*+N.Pl.Gen.PxSg1.

Thus, the development of a working translation system will provide grounds for expanding research on the dialects and languages of the closely related Mordvin languages, and, we hope, it will indicate research objects for a fuller coverage of language description.

Albeit, morphological comparisons have been proposed for the alignment of otherwise diverse morpho-syntactic readings, but there are still matters of unknown correlations (see examples 1–4). Problems in establishing parallels between morphological categories in the source and target languages can be found in nouns, numerals and verbs [c.f. Rueter, Partanen & Ponomareva, 2020].

The nominal declension affords an intricate interface between Erzya and Moksha. Where Erzya has definite marking for 9–10 singular cases and 12–13 plural cases [Rueter 2010: 99–100], Moksha only attests to three cases in the definite declensions. Useful contemplation of this asymmetry can be detected, e.g. Keresztes (2011: 68; Zaicz 1990) draws a parallel between analytic and synthetic expression in Erzya noun morphology and analytic expression in Moksha. In communication, however, one is not interested in the etymological parallels that can be drawn between constructions, instead there is a need for semantic parallels. Thus the Keresztes solution, where five prolative forms in Erzya can be equated with three in Moksha must be adjusted. First, whereas Erzya uses the elative case in both nouns and postpositions, Moksha uses an elative form in the indefinite nominal declension but an ablative form in the postpostion. Second, we do not know whether the Erzya and Moksha categories of definiteness are parallel, as comparative studies of definiteness in Uralic language pays little attention to the Mordvin languages where definite/determinate marking are distinguished from possessive marking (see [Tikhonova 1966]), and insufficient data is present for comparison of Mordvin and Mordvin speakers of Russian, such as Avvakum [cf. Abramov V. K. 2003: 97; Yurayong 2020: 100–101, 105–106]. And third, we do not know what the exact correlations are between the Erzya five and the Moksha three (see examples 3–4).

(3a)
(myv) *kudo-sto*
house_N-SP.Ela.Indef

(3b)
(myv) *kudo-sto-ńť*
house_N-Ela-Sg.Def

(3c)
(myv) *kudo-ńť* *ejste*
house_N-Sg.Gen.Def Po.Ela

(3d)
(myv) *kudo-ťńe-ste*
house_N-Pl.Def-Ela

(3e)
(myv) kudo-t.ńe.ń ejste
house_N-Pl.Def-Gen Po.Ela

(4a)
(mdf) *kud-sta*
house_N-SP.Ela.Indef

(4c)
(mdf) *kud-ť* *ezda*
house_N-Sg.Gen.Def Po.Abl

(4d)
(mdf) kut-t.ńe.ń ezda
house_N-Pl.Gen.Def Po.Abl

While the five Erzya elative forms might be equated to the Moksha one elative plus two ablative forms, it will be simpler to address the prolative case (see examples 5–6), which, in addition to retaining one case in both languages, tends to show an alignment between the definite Erzya synthetic representation and the definite Moksha analytic construction; use of the Erzya analytic construction in the singular definite/determinate genitive is only attested 109 times in the ERME Sura-Sjatko corpus 1956–2000, whereas the synthetic construction can be found well over 5500 times. Hence, here it will suffice for initial work to align the pairs (5a=6a), (5b=6b) and (5d=6c). From a linguistic perspective, it would be interesting to see whether variation in Erzya prolative marking is simply a matter of spatial-temporal complementary distribution pertinent to Mordvin dialect studies [cf. Rueter 2016; 2020], or is there something more to it, as might be noted of at least Komi-Zyrian [cf . Nekrasova, 2019].

(5a)
(myv) *kudo-va*
house_N-SP.Prl.Indef
(5b)
(myv) *kudo-va-ńt́*
house_N-Prl-Sg.Def
(5c)
(myv) *kudo-ńt́* *ezga*
house_N-Sg.Gen.Def Po.Prl
(5d)
(myv) *kudo-t́ńe-va*
house_N-Pl.Def-Prl
(5e)
(myv) kudo-t.ńe.ń ezga
house_N-Pl.Def-Gen Po.Prl
(6a)
(mdf) *kud-ga*
house_N-SP.Prl.Indef
(6c)
(mdf) *kud-t́* *ezga*
house_N-Sg.Gen.Def Po.Prl
(6e)
(mdf) kut-t.ńe.ń ezga
house_N-Pl.Gen.Def Po.Prl

The morphological categories of case, number, definiteness and person are shared with other nominals. In a word, when the NP head in contextually retrievable positions is elided, the subsequently promoted NP head automatically becomes the locus of NP category marking. This said, we can briefly examine the intricacies of Mordvin verbs.

From a Uralic studies and perhaps even a typological perspective, Mordvin verb morphology is known for its subject and subject-object[11] conjugation paradigms as well as a relatively elaborate set of moods. Within the language group, however, the most blatant places of asymmetry can be found in the subject-object conjugation and the second preterite (see [Rueter & Tyres, 2018]). In Erzya, subject-object marking tends to be limited to marking definite objects with stative verbs, on the one hand, and the addition of a total object stipulation when used with non-stative verbs, on the other. Moksha may also deviate from the total object stipulation especially in the case of frequentative verbs (see [Bartens, 1999: 125, 175]).

If we examine the finite verbal paradigms of the two literary Mordvin languages more extensively, we will note that 48% of 196 combinatory functions in the finite verbal paradigms involve either unambiguity or shared ambiguity (see [Rueter 2016]; in press: Linguistic distance between Erzya and Moksha, dependent morphology). An instance of unambiguous would, for example, involve a 1Sg subject and a 2Sg object. Shared ambiguity, however, might involve an entire series with a default 2nd person object in combination with the number category value plural applied to one or both of the core arguments of the verb. This means there is a lot of research required for disambiguation between languages. In this paper, we will confine ourselves to the use of the non-ambiguous form: indicative 3Sg Prt1, which will appear in our simple translation sentence (see 7, below).

First Coding Challenges

In the Apertium Erzya-Moksha project we have started with pre-defined morphology and lexica (see [Rueter, Hämäläinen & Partanen 2020]). The pre-defined morphology, consisting of finite-state descriptions, and lexica provide elementary components for the development of Apertium shallow-transfer machine translation for the two Mordvin literary languages. Consistent with Mordvin study practices, descriptions of the two literary languages are often drafted in a manner where parallel and mutual features are aligned, and deviation from this unity is less regular. This practice can be utilized here for purposes of delimiting an example sentence as well as illustrating a few dimensions where shallow-transfer helps highlight morphological, lexical and syntactic features of the Erzya and Moksha siblings.

11 Unlike the Amazon language Apurinã, where the verbs can take subject marking, object marking or a combination of the two, Mordvin verbs take either subject marking or a combination of subject and object marking [cf. Facundes 2000: 269–343, Freitas 2017: 84–92].

For purposes of illustration, we will focus on a three-word example sentence that can be translated from Erzya to Moksha and back again using the shallow-transfer concept. Here we have chosen a sentence meaning ‘The cat walked around in the house.’ (see 7.1–7.2).

(7.1)

Psaka-ś	*jaka-ś*	*kudo-va-ńť*
cat_N-Sg.Nom.Def	walk_V-Ind.Prt1.ScSg3	house_N-Prl-Sg.Def

The three words of the Erzya sentence in (7.1) consist of two nouns in the definite/determinate declension and an intransitive verb third person singular in the first preterite. When translated into Moksha, a typical rendering would consist of two nouns with the same semantic content, but the word for ‘cat’ would have two distinct etymological backgrounds, whereas the verb ‘to walk’ in *jakams* is of mutual background. On a syntactic level, we can also observe the alignment of the Erzya synthetic construction *kudo-va-ńť* house_N-Prl-Sg.Def ‘around in the house’ with the Moksha analytic construction *kud-ť ezga* consisting of a noun house_N-Sg.Gen.Def and a subsequent adposition Po.Prl ‘around in the house’.

(7.2)

Kato-ś	*jaka-ś*	*kud-ť*	*ezga*
cat_N-Sg.Nom.Def	walk_V-Ind.Prt1.ScSg3	house_N-Sg.Gen.Def	Po.Prl

The simple sentence ‘The cat walked around in the house.’ in *psakaś jakaś kudovańť* = Псакась якась кудовантъ.

Can be called in the Apertium tool with:

echo Псакась якась кудовантъ. | apertium -d. myv-mdf-biltrans

```
^Псака<n><sg><nom><def><@X>/Ката<n><sg><nom><def><@X>$
 ^якамс<vblex><indic><pret1><s_sg3><@X>/@якамс<vblex><indic><pret1><s_sg3><@X>$
 ^кудо<n><sg><prl><def><@X>/куд<n><sg><prl><def><@X>$
^..<sent>/..<sent>$
```

Figure 1. Bilingual transfer Erzya-to-Moksha ‘The cat walks around in the house.’

The first lemma in the upper left is *psaka* ‘cat’, which is upper case, since it comes at the beginning of the sentence. ‘Cat’ is followed by the tags <n> for noun, <sg> for singular, <nom> for nominative and <def> for definite. This is also the order the analyzer has been set up to produce in nearly all languages in the Giellatekno infrastructure, where the morphological description is made. The third lemma, *kudo* ‘house’, with its four tags only deviates from that of the first in the third tag from the left, <prl> for prolative. For this reason, part-of-speech, and additional grammatical categories can be directly transferred from the source language to the target language. If necessary, the tags can then be further dealt with to match what is needed for the target language.

The second lemma on the left is *jakams* ‘to walk’, which is followed by the tags <vblex> for regular verb, <indic> for indicative, <pret1> for preterite 1 and <s_sg3> for third person singular, subject conjugation. The seemingly superfluous declarations of part-of-speech and morphological zero values such as indicative are for purposes of symmetry needed in the Apertium toolkit. The computer only possesses the information we have, so we give it that information here. One item that has been left out of the verbal analysis is the category of transitivity, which, although important for some languages in disambiguation, is only one of the many matters to be considered when dealing with verbal argument structure.

The Erzya noun *psaka* ‘cat’ can readily be aligned with Moksha noun *kata* through use of the bilingual dictionary, apertium-myv-mdf.myv-mdf.dix. The same can be achieved for the verb *jakams* ‘to walk’. While the noun *kudo* ‘house’ can also be dealt with using the same .dix file. Conversion in structures, however, takes us back to alignment hypotheses,

such as those noted above [Keresztes 2011], involving combinations of the morphological categories definiteness, case and number with distinct reflexes in Erzya and Moksha.

Erzya, on the one hand, has two representations of the combination singular + prolative + definite, while Moksha has but one, on the other. The most common representation in Erzya is a synthetic structure, which with the word *kudo* 'house' can be shown as *kudo-va-ńť* with the gloss house_N-Prl-Def.Sg 'around in the house'. The Erzya morphological structure, consisting of one word form, must then be transformed to a two-word structure for Moksha, where the word for house is generated in the definite genitive singular followed by a postposition representing the prolative case (see analysis in example 8.1 and generation in 8.2–3).

8.1 (Erzya) *kudovańť* => *kudo*+N+Sg+Prl+Def
8.2 (Moksha) *kud*+N+Sg+Gen+Def PLUS Po+Prl
8.3 (Moksha) => *kudť ezga*

This transfer from the synthetic vs analytic chunks can be dealt with using a special rule in the apertium-myv-mdf.myv-mdf.t1x file (see Figure 2.). This rule transfers two nominal categories: number and definiteness from the Erzya stem to the Moksha stem automatically, whereas the prolative case marker at this point is hard-coded into the prolative postposition *ezga*.

```
<rule comment="noun in definite prolative: is definite genitive noun with ezga">
  <pattern>
    <pattern-item n="n.prl.def"/>
  </pattern>
  <action>
    <out>
      <chunk name="PrlDef" case="caseFirstWord">
        <tags>
          <tag><lit-tag v="SN"/></tag>
        </tags>
        <lu>
          <clip pos="1" side="tl" part="lem"/> <!-- sourceL lemma -> targetL lemma -->
          <clip pos="1" side="tl" part="a_nom"/> <!-- Part-of-Speech -->
          <clip pos="1" side="tl" part="a_number"/> <!-- Number category -->
          <lit-tag v="gen"/> <!-- Target language case requirement -->
          <clip pos="1" side="tl" part="a_definiteness"/> <!-- Category of definiteness -->
        </lu>
        <b/>
        <lu>
          <lit v="əзra"/>
          <lit-tag v="adp"/>
          <lit-tag v="prl"/>
        </lu>
      </chunk>
    </out>
  </action>
</rule>
```

Figure 2. Transfer rule: Erzya N.Sg.Prl.Def => Moksha N.Sg.Gen.Def Po.Prl

The pattern-item, corresponding to a definite (*def*) Erzya noun in the prolative case (*prl*) brings about an action with two lexical units (*lu*). The first lexical unit, Erzya *kudo* 'house' is transferred by lemma (*lem*) to the Moksha equivalent *kud*. The Moksha noun *kud* 'house' then inherits, as it were, the categories of number (*a_number*) and definiteness (*a_definiteness*), and the genitive (*gen*) case is added as a target language case requirement for the postpositional complement. Then a prolative postposition is introduced with the second lexical unit.

With everything in place and no other ambiguity the command-line provides us with an accurate output to Moksha:

echo ”Псакась якась кудованть” | apertium -d . myv-mdf
Катось якась кудть эзга
‘The cat walked around in the house’

Future Work and Conclusions

We have described some of the current work conducted on applying the Apertium shallow-transfer machine translation toolkit to the Erzya and Moksha languages. In addition, we have shed some light into the future directions that this work can take and how it can be of a mutual benefit for linguistics and language technology research.

As we have described in this paper, some of the rule-based methods in use in Apertium can also be modeled with neural networks to a degree. Therefore it is an important question to be answered in the future to what degree machine learning based modules can be plugged in to the Apertium pipeline. Instead of having a machine translation solution that relies entirely on rules or neural machine translation, it is more fruitful to look at the intersection of both approaches.

Nevertheless, in order to gain the maximum benefit from a shallow-transfer based approach, it is important to focus on building more parallel lexica. Preferably in such a fashion, that the lexica are multilingual rather than bilingual as it makes it possible to have multiple translation directions.

Important work has been done with open-source lexica in the Finnotka project (Ferenczi et al 2018), where limited open-source texts have been utilized for producing initial parallel lexicon.

The Apertium Project has been able to utilize Google Summer of Codes (GSoC) projects to advance a large variety of necessary code. One outcome has been a month’s worth of Udmurt-Komi-Zyrian lexica for the translation toolkit and this will merit further investigation into use of non-speaker programming for lexical enhancement.

Harmonization in the morphological tagging practices, the transferability of morpho-semantic categories and successful alignment of syntactic dependencies between languages all require continuous constructive discourse and planning. As of November 15th 2020, there are already 11 Uralic languages represented in the Universal Dependencies project: Estonian, Finnish, Karelian, Livvi-Karelian; North Sami, Skolt Sami; Erzya, Moksha; Komi-Permyak, Komi-Zyrian; Hungarian (see [Zeman, Daniel et al 2020]).

Grouping language projects at Apertium into clusters of closely related languages means that we can follow the lead of work with the Romance languages at Apertium, as well as work with Balto-Finnic language. At present there are 14 Uralic languages involved at varied levels of progress in work with the Apertium toolkit: Estonian, Finnish, Olonets-Karelian, Karelian; Hungarian; Komi-Zyrian, Komi-Permyak, Udmurt; Hill Mari, Meadow/Eastern Mari; Inari Sami, Northern Sami; Erzya, Moksha. The well over 20 apertium language pairs including at least one Uralic language provide further indication of reuse and new outcomes for fieldwork vocabulary, syntax and other amassed in Finno-Ugric and Uralic studies over the past couple of centuries. These language pairs (fin-deu, fin-eng, fin-est, fin-fkv, fin-hun, fin-krl, fin-nor, fin-olo, fin-rus, fin-udm, kpv-fin, kpv-mhr, liv-fin, mrj-fin, myv-fin, olo-fin, sme-fin, udm-kpv, udm-rus, tur-fin, etc.) may provide for parallel development and testing for shared translation projects. The presence of non-related language pairs is also beneficial, as they may provide insight into transfer rule construction for phenomena not encountered elsewhere.

Acknowledgements

The Erzya-Moksha translation pair will best learn from the languages with several translation targets or sources. We have received an abundance of assistance from Ph.D. T. Trosterud for experience with translation between Uralic languages at Giellatekno, Ph.D.T. Pirinen and his

work with the Finnish-German pair, Ph.D. F. Tyers and Ph.D. J. Washington for their interest and insights on language category transfer as well as T. Didriksen and K. Unhammer for making sure the Apertium infrastructure is working right[12].

References and sources

Abramov, V. K. Аввакум [Avvakum] // Мордовия, Энциклопедия в двух томах, Том 1 А–М. Главнай пелакцирееая коллегия: Главный редактор А. И. Сухарев, Заместитель главного редактора В. А. Юрченков, Ответственный секретарь П. П. Кузнецов, И. М. Арсентьев, Ш. И. Ахметов, Э. А. Боксимер, Н. В. Бурмистров, архиепископ Варсонофий, Г. Г. Вдовин, С. М. Вдовин, В. Д. Волков, Е. В. Воскресенский, Н. И. Еналеев, Н. В. Заварюхин, Г. М. Измалкин, И. А. Калинкни, Б. Ф. Кевбрин, В. А. Кечкин, А. М. Кокинов, В. А. Кокорев, В. В. Конаков, Н. С. Куратов, В. В. Литюшкин, А. С. Лузгин, Н. П. Макаркин, В. М. Макушкин, Н. Ф. Мокшин, Е. Г. Осовский, Н. В. Петрушкин, А. О. Пиксаев, В. А. Попков, А. М. Пыков, В. П. Селяев, А. И. Солдатов, С. Ф. Сорокин, Б. И. Стародубов, Т. В. Тюрина, П. В. Шичкин, Д. В. Цыганкин, В. Д. Черкасов. – Саранск: Мордовское книжное издательство, 2003. P. 97.

Alnajjar, K., Hämäläinen, M., & Rueter, J. On Editing Dictionaries for Uralic Languages in an Online Environment // Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages. The Association for Computational Linguistics. 2020. Pp. 26–30.

Alnajjar, K., Hämäläinen, M., Rueter, J., & Partanen, N. Ve'rdd. Narrowing the Gap between Paper Dictionaries, Low-Resource NLP and Community Involvement // Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations. 2020.

Bartens, R. Mordvalaiskielten rakenne ja kehitys. Suomalais-Ugrilaisen Seuran Toimituksia 232. Helsinki. 1999.

Corbí Bellot, A. M., Forcada, M. L., Ortiz Rojas, S., Pérez-Ortiz, J. A., Ramírez Sánchez, G., Sánchez-Martínez, F., Alegría Loinaz, I., Mayor Martínez, A., & Sarasola Gabiola, K. An open-source shallow-transfer machine translation engine for the Romance languages of Spain // Proceedings of the 10th Conference of the European Association for Machine Translation. 2005

EKM – Д. В. Цыганкин (отв. ред.) и др., Эрзянь кель. Морфология. – Саранск: Красный Октябрь, 2000.

EKS – Эрзянь кель. Синтаксис : тонавтнемапель / Н. А. Агафонова, Р. А. Алёшкина, Г. Ф. Беспалова [ды лият] ; анокстазь Д. В. Цыганкинэнь ветямонзо ало. Саранск: Изд-во Мордов. ун-та. 2011. 208 с.

Ens, J., Hämäläinen, M., Rueter, J., & Pasquier, P. Morphosyntactic Disambiguation in an Endangered Language Setting. In M. Hartmann, & B. Plank (Eds.), 22nd Nordic Conference on Computational Linguistics (NoDaLiDa): Proceedings of the Conference (Linköping Electronic Conference Proceedings; No. 167), (NEALT Proceedings Series; No. 42). Linköping University Electronic Press. 2019. Pp. 345–349.

Erina, O. & Rueter, J. Shoksha Kolkhoznikin' val'gij 1932–1933 (Version v1.0) https://zenodo.org/badge/120288368.svg (Collection of proofread texts and originals). 2018.

ERME: Erzya and Moksha Extended Corpora. Rueter, Jack & Olga Yerina (eds., compilers.). (Partially accessible: http://gtweb.uit.no/u_korp/?mode=myv, https://korp.csc.fi/korp/?mode=other_languages#?lang=fi&stats_reduce=word&cqp=%5B%5D&corpus=erme_myv).

Evsev'ev, M. E. – Евсевьев, М. Е. Основы Мордовской Грамматики – Эрзянь грамматика. 1928/29.

Facundes, Sidney Da Silva. The Language of the Apurinã People of Brazil (Maipure/Arawak). (A Dissertation submitted to the Faculty of the Graduate School of State University of New York at Buffalo in partial fulfillment of the requirements for the degree of Doctor of Philosophy Department of Linguistics). 2000.

Ferenczi, Zsanett; Mittelholcz, Iván; Simon, Eszter. Automatic Generation of Wiktionary Entries for Finno-Ugric Minority Languages // Tommi A. Pirinen, Michael Riessler, Jack Rueter, Trond Trosterud, Francis M. Tyers (eds.): Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages. Association for Computational Linguistics. 2018. Pp. 39–50.

12 Here are a couple of irc channels on the chat.freenode.net where we have been able to get almost instant feedback #apertium (Apertium development), #hfst (Helsinki Finite-State Technology for Giellatekno infrastructural questions), #_u-dep (discussions pertaining to Universal Dependencies projects)

*Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ram*írez-Sánchez, G. & *Tyers, F. M.* Apertium: a free/open-source platform for rule-based machine translation. Machine translation. 2011. 25(2). Pp. 127–144.

Freitas, Marília Fernanda Pereira De. A POSSE EM APURINÃ: descrição de construções atributivas e predicativas em comparação com outras línguas Aruák. (Tese apresentada ao Programa de Pós-graduação em Letras da Universidade Federal do Pará – Doutorado acadêmico em Letras – estudos Linguísticos, como requisito para obtenção do título de Doutor. Orientador: Prof. Dr. Sidney da Silva Facundes. Coorientadora: Profª. Drª. Patience Epps). 2017. Pp. 84–92.

Garrett, A. Online dictionaries for language revitalization. In The Routledge Handbook of Language Revitalization. UC Berkeley. 2018.

Gheno, Danilo "Mordwinisch" oder "Mokschanisch und Erzanisch"? In: Zaicz (ed), Zur Frage der uralischen Schriftsprachen. Linguistica. Series A: Studia et dissertationes, 17. Budapest, 1995. Pp. 57–61.

Hamari, Arja. The negation of stative relation clauses in the Mordvin languages. MSFOu 254. Helsinki, Suomalais-Ugrilainen Seura. 2007.

Hämäläinen, M., & Alnajjar, K. A Template Based Approach for Training NMT for Low-Resource Uralic Languages - A Pilot with Finnish // ACAI 2019: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. 2019a. Pp. 520–525. ACM.

Hämäläinen, M., & Alnajjar, K. Let's FACE it: Finnish Poetry Generation with Aesthetics and Framing // K. V. Deemter, C. Lin, & H. Takamura (Eds.), 12th International Conference on Natural Language Generation: Proceedings of the Conference. 2019b. Pp. 290–300.

Hämäläinen, M., & Rueter, J. (). Development of an Open Source Natural Language Generation Tool for Finnish // Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages. 2018. Pp. 51–58. The Association for Computational Linguistics.

Hämäläinen, M., Tarvainen, L. L., & Rueter, J. Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages // N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018. Pp. 862–867.

Hämäläinen, M., & Wiechetek, L. Morphological Disambiguation of South Sámi with FSTs and Neural Networks // Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020). 2020 Pp. 36–40. European Language Resources Association (ELRA).

Hämäläinen, M. UralicNLP: An NLP Library for Uralic Languages. Journal of open source software, 2019. 4(37). [1345]. https://doi.org/10.21105/joss.01345.

Keresztes, L. Bezetés a mordvin nyelvészetbe. Debreceni Egyetemi Kiadó.

Kocmi, T., & Bojar, O. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation: Research Papers 2011. 2018. Pp. 244–252.

Luutonen, Jorma et al. Lexica Societatis Fenno-Ugricae XXXI: I. Electronic Word Lists: Mari, Mordvin and Udmurt with SFOu WordListTool 1.3 for PC and Mac. Finno-Ugrian Society. 2007.

MKM – Мокшень кяль. Морфология: Вузонь мокшень и финно-угрань отделениянь тонафнихненди учебник / Серматф-тиф Н. С. Алямкинонь кядяла. Саранск: Тип. «Крас. Окт.», 2000. 236 с.

MKS – Мокшень кяль. Синтаксис: Учебник. / аноклаф-тиф Н. С. Алямкинонь и О. Е. Поляковонь профессорхнень вятемаснон ала. Саранск: Изд-во Мордов. ун-та, 2008. 200 с.

Moshagen, Sjur; Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages // *The LREC 2014 Workshop "CCURL 2014- Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era"*. 2014. Pp. 71–77.

Nekrasova, G. A. – Некрасова, Г. А. «Конкуренция пролативных суффиксов в коми языке: Опыт корпусного исследования." Актуальные вопросы коми и пермского языкознания. Труды Института языка, литературы и истории Коми НЦ УрО РАН. 2019. Выпуск 77. Pp. 53–63.

Ngo, T. V., Nguyen, P. T. , Ha, T. L., Dinh, & K.Q., Nguyen, L. M. Improving Multilingual Neural Machine Translation For Low-Resource Languages: French, English - Vietnamese // Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages. 2020. Pp. 55–61.

Paasonen, Heikki – Lexica Societatis Fenno-Ugricae XXIII,1-4: H. Paasonens Mordwinisches Wörterbuch zusammengestellt von Kaino Heikkilä, Bände I-IV, unter mitarbeit von Hans-Hermann Bartens, Aleksandr Feoktistow und Grigori Jermuschkin, bearbeitet und herausgegeben von Martti Kahla. 1990-1996. (XXXI-LXXXVI). Suomalais-Ugrilainen Seura.

Partanen, Niko & Rueter, Jack. Survey of Uralic Universal Dependencies development // Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019). 2019. Pp. 78–86.

Partanen, N., Blokland, R., Lim, K., Poibeau, T., & Rießler, M. The First Komi-Zyrian Universal Dependencies Treebanks. In Second Workshop on Universal Dependencies (UDW 2018). Brussels, 2018. Pp. 126–132.

Pirinen, T. A. Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. SKY Journal of Linguistics. 2015. 28.

Pirinen, T. A. Apertium-fin-eng–Rule-based Shallow Machine Translation for WMT 2019 Shared Task // Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019. Pp. 335–341.

Reynolds, R. J. Russian natural language processing for computer-assisted language learning. UiT The Arctic University of Norway Doctoral Dissertation. 2016.

Rueter, J. & Hämäläinen, M. FST Morphology for the Endangered Skolt Sami Language // D. Beermann, L. Besacier, S. Sakti, & C. Soria (Eds.), Proceedings of the 1st Joint SLTU and CCURL Workshop. 2020. Pp. 250–257. European Language Resources Association (ELRA).

Rueter, J. M., & Hämäläinen, M. Synchronized Mediawiki based analyzer dictionary development // 3rd International Workshop for Computational Linguistics of Uralic Languages Proceedings of the Workshop. Association for Computational Linguistics. 2017

Rueter, J., Hämäläinen, M., & Partanen, N. Open-Source Morphology for Endangered Mordvinic Languages // Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS). 2020. pp. 94–100. The Association for Computational Linguistics.

Rueter, J., Partanen, N., & Ponomareva, L. On the questions in developing computational infrastructure for Komi-Permyak // Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages. 2020. Pp. 15–25.

Rueter, Jack and Tyers, Francis. Towards an open-source universal-dependency treebank for Erzya // Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages (IWCLUL2018). 2018. Pp. 106–118.

Rueter, Jack. Adnominal Person in the Morphological System of Erzya. Suomalais-Ugrilaisen Seuran Toimituksia 261. Helsinki, 2010. Suomalais-Ugrilainen Seura.

Rueter, Jack. The Livonian-Estonian-LatvianDictionary as a threshold to the era of language technological applications. Eesti ja soome-ugrikeeleteaduse ajakiri. 2014. 5(1). Pp. 251–259.

Rueter, Jack. Towards a systematic characterization of dialect variation in the Erzya speaking world, Isoglosses and their reflexes attested in and around the Dubyonki Raion // Ksenia Shagal with Heini Arjava (eds): Mordvin languages in the field. Uralica Helsingiensia 10. 2016.

Rueter, Jack (in press). Linguistic distance between Erzya and Moksha, dependent morphology.

Rueter, Jack. Корпус национальных мордовских языков: принципы разработки и перспективы функционирования/действия [A Corpus of National Mordvin Languages: principles of development and perspectives of functionality]. Финно-угорские народы в контексте формирования общероссийской гражданской идентичности и меняющейся окружающей среды.: материалы междунар. науч. конф. г. Саранск, 8–9 октября 2020 г. / [отв. ред. чл.-корр. РАН Н. М. Ансентьев]. Саранск: Изд. центр ист.-социал ин-та, 2020. Pp. 118–127.

Salimzyanov, I., Washington, J., & Tyers, F. A free/open-source Kazakh-Tatar machine translation system. Machine Translation Summit XIV. 2013 Pp. 175–182.

Snoek, C., Thunder, D., Loo, K., Arppe, A., Lachler, J., Moshagen, S., & Trosterud, T. Modeling the noun morphology of Plains Cree // Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages. 2014. Pp. 34–42.

Sundetova, A., Karibayeva, A., & Tukeyev, U. Structural Transfer Rules for Kazakh-to-English Machine Translation in The Free/Open-Source Platform Apertium. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi. 2014. 7(2). Pp. 48–53.

Tikhonova, T. M. (). = Tichonova, T. M. Expression of definiteness and indefiniteness of the direct object in the Mordvin languages // Советское финно-угроведение. 1966. (4), Pp. 241–245.

Trosterud, S.R., Trosterud, T., Räisänen, A.K., Niiranen, L., Haavisto, M., & Maliniemi, K. A morphological analyser for Kven. Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages. W17–0608. 2017. Pp. 76–88.

Trosterud, T. A constraint grammar for Faroese. Constraint Grammar and robust parsing. 2009. 1.

Tsypkaikina V. P. – Цыпкайкина, В. П.: Глаголось. Эрзянь кель, морфология, Pp. 187–190. Эрзянь кель, морфемика, валонь теевема ды морфология: Вузонь эрзянь ды финнэнь отделениянь тонавтницятнень туртов / Редколлегиясь: Д. В. Цыганкин (отв. ред., Н. А. Агафонова, М. Д. Имайкина ды лият. – Саранск: Тип. «Крас. Окт.». 2000. 280 с.

Tyers, F. M., Washington, J. N., Salimzyanov, I., & Batalov, R. A prototype machine translation system for Tatar and Bashkir based on free/open-source components. In First Workshop on Language Resources and Technologies for Turkic Languages. 2012.

Uí Dhonnchadha, E., & van Genabith, J. A Part-of-Speech tagger for Irish using finite state morphology and constraint grammar disambiguation. 2006.

Yurayong, Chingduang. Postposed demonstratives in Finnic and North Russian dialects. Department of Finnish, Finno-Ugrian and Scandinavian Studies. Faculty of Arts, University of Helsinki. (Doctoral Dissertation). 2020.

Zaicz, Gábor. Beitrag zur Typologie und Statistik des erza- und mokscha-mordvinischen, (eine vergleichende Untersuchung) // *A/5. Volgai* és *permi nyelészet*. Budapest/Uppsala, 1990. Pp. 7–13.

Zaicz, Gábor. Сколько языков нужно ерзе и мокше? // Zaicz (ed), Zur Frage der uralischen Schriftsprachen. Linguistica, Series A, Studia et dissertationes, 17. Budapest. 1995. Pp. 41–46.

Zaicz, Gábor. Mordva // Abondolo (ed.), The Uralic Languages. London and New York, Routledge. 1998. Pp. 184–218.

Zeman, Daniel; *Nivre, Joakim; Abrams, Mitchell; et al.* Universal Dependencies 2.7, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3424. 2020.

Джек Рютер
PhD, исследователь
Финляндия, Хельсинки, Хельсинкский университет

Мика Хямялайнен
PhD, исследователь
Финляндия, Хельсинки, Хельсинкский университет

ПРЕДПОСЫЛКИ ДЛЯ МАШИННОГО ПЕРЕВОДА МОРДОВСКИХ ЯЗЫКОВ НА ОСНОВЕ ПОВЕРХНОСТНОГО ТРАНСФЕРА

This paper presents the current lexical, morphological, syntactic and rule-based machine translation work for Erzya and Moksha that can and should be used in the development of a roadmap for Mordvin linguistic research. We seek to illustrate and outline initial problem types to be encountered in the construction of an Apertium-based shallow-transfer machine translation system for the Mordvin language forms. We indicate reference points within Mordvin Studies and other parts of Uralic studies, as a point of departure for outlining a linguistic studies with a means for measuring its own progress and developing a roadmap for further studies.

Keywords: *Erzya, Moksha, Uralic, Shallow-transfer machine translation, Measurable language research, Measurable language distance, Finite-State Morphology, Universal Dependencies.*